POLVO DE ESTRELLAS: SESGO DE GÉNERO Y ESCALAS DE EVALUACIÓN EN LA CRÍTICA DE CINE

Maria Cubel
City St George's University London

Santiago Sanchez-Pages
King's College London

Resumen

Este artículo analiza cómo la arquitectura de las evaluaciones influye en el sesgo de género en la crítica cinematográfica. Usamos más de 4.000 reseñas publicadas entre 2021 y 2025 en cuatro medios españoles, dos con escalas de estrellas enteras y dos con medias estrellas. Estimamos modelos de efectos aleatorios y fijos que predicen la puntuación media y la probabilidad de obtener la máxima calificación, controlando por características de las películas y el género del crítico. Encontramos que las películas dirigidas por mujeres reciben valoraciones inferiores, sobre todo por críticos varones, y que esta penalización se atenúa en escalas más finas.

Palabras clave: género, sesgos, evaluación, crítica cinematográfica.

Abstract

This article examines how the architecture of evaluation scales influences gender bias in film criticism. We use more than 4,000 reviews published between 2021 and 2025 in four Spanish outlets, two employing full-star scales and two using half-star ratings. We estimate random-effects and film fixed-effects models that predict both the average rating and the probability of receiving the maximum score, controlling for film characteristics and critic gender. We find that films directed by women receive lower evaluations, especially from male critics, and that this penalty diminishes when more granular scales are available.

Keywords: gender, bias, evaluation, film criticism.

JEL classification: J16, Z11.

I. INTRODUCCIÓN

Las escalas numéricas de evaluación se utilizan en múltiples contextos. Su objetivo declarado es medir de forma comparable y objetiva el desempeño individual y facilitar la asignación de contratos, remuneraciones o promociones. Este tipo de evaluaciones está presente en la docencia universitaria, la formación médica, la selección de personal y los mercados laborales digitales. Desde hace décadas, disciplinas como la psicología y la pedagogía las han empleado también para estandarizar la medición de habilidades y actitudes.

Sin embargo, la evidencia acumulada muestra que estos instrumentos numéricos están lejos de ser neutrales. Los sesgos cognitivos y sociales de quienes evalúan generan diferencias sistemáticas en calificaciones. En concreto, las mujeres tienden a ser evaluadas con mayor severidad. Sus logros se reconocen en menor medida, su autoridad se cuestiona con más frecuencia y es menos probable que sean descritas como "brillantes" o "excepcionales", categorías reservadas de manera desproporcionada a los hombres (Leslie *et al*., 2015). Numerosos estudios documentan la presencia de estos sesgos de género en distintos ámbitos: las mujeres reciben sistemáticamente peores evaluaciones docentes (Boring, 2017; Mengel *et al*., 2019; Fitzpatrick *et al*., 2021); afrontan estándares más exigentes para publicar en revistas académicas y acceder a congresos (Card *et al*., 2020, 2022; Hengel, 2022; Hospido y Sanz, 2021; Samahita, 2024), y sus propuestas de financiación son evaluadas menos

favorablemente (Witteman *et al*., 2019). Estos sesgos tienen efectos acumulativos: ralentizan la progresión profesional, reducen la remuneración y perpetúan la infrarrepresentación femenina en ciertos campos (Alan *et al*., 2018; Dupas *et al*., 2021; Eberhardt *et al*., 2023).

La investigación reciente ha comenzado a atender al papel que juega la arquitectura de la evaluación en la expresión de estos sesgos. Bajo este concepto se incluyen características formales del diseño de las escalas de evaluación: el número de categorías, su etiquetado o la forma de presentar los resultados. Estas dimensiones, aunque aparentemente técnicas, inciden en la manifestación de sesgos. Por ejemplo, Rivera y Tilcsik (2019) observaron que pasar de una escala de 10 a una de 6 puntos elimina las brechas de género en evaluaciones docentes, especialmente en disciplinas masculinizadas. Botelho *et al*. (2025) y Rivera (2025) encuentran que la dicotomización de escalas multipunto en mercados laborales digitales elimina diferencias raciales y salariales. Estos resultados sugieren que el diseño de las escalas afecta tanto a su fiabilidad y validez como al grado en que los sesgos de género y otras formas de discriminación se reflejan en las evaluaciones.

En el presente artículo estudiamos esta cuestión en el ámbito de la crítica cinematográfica. En las industrias culturales, donde los juicios de valor son necesariamente subjetivos, la crítica ejerce una influencia decisiva sobre trayectorias y reputaciones. El uso de estrellas, puntuaciones sobre 10 o sobre 100 y, más recientemente, los algoritmos de agregación como Rotten Tomatoes o Metacritic han convertido las evaluaciones numéricas en un lenguaje universal para sintetizar juicios de valor sobre el cine.

Las industrias creativas no son ajenas a los patrones de sesgo de género existentes en entornos técnicos y corporativos. En el caso del cine, la desigualdad comienza ya en la producción: las mujeres ocupan apenas una cuarta parte de los roles creativos clave en la mayoría de las cinematografías nacionales (Berry *et al*., 2022; Loist *et al*., 2024). En el ámbito crítico, las mujeres siguen siendo una minoría pese a avances recientes (VIDA Foundation, 2020). Además, la figura del "autor" y del "genio creativo" continúa asociada de manera predominante a los hombres (Nochlin, 1972; Proudfoot *et al*., 2015). Este sesgo de género impregna también la recepción crítica. Como documentamos en un trabajo anterior (Cubel y Sanchez-Pages, 2025), las películas dirigidas por mujeres reciben puntuaciones inferiores, en especial por parte de los críticos varones, incluso controlando por otros indicadores de calidad. La pregunta de investigación que guía este artículo es si esa penalización puede verse modulada por la arquitectura de la evaluación. Esta cuestión es relevante porque la crítica, en tanto que instancia de legitimación cultural, puede contribuir a reforzar o a mitigar las desigualdades existentes.

Para ello, analizamos una muestra de más de 4.000 reseñas publicadas entre 2021 y 2025 en cuatro medios digitales españoles. Dos de ellos utilizan una escala de estrellas enteras (5 puntos) para puntuar las películas que reseñan, mientras que los otros dos admiten medias estrellas (9 puntos). Esto nos permite examinar cómo la granularidad de la escala modula la expresión del sesgo de género. Combinamos información a nivel de película y de crítico, incorporando datos sobre las características de los filmes, su recepción y la identidad del revisor. Estimamos modelos de efectos aleatorios y especificaciones con efectos fijos de película para comparar valoraciones dentro de un mismo título, analizando tanto la puntuación media como la probabilidad de recibir la máxima calificación en función del género del equipo directoral y de la persona que escribe la reseña.

Nuestros resultados confirman en parte la literatura previa que muestra que la arquitectura de la escala altera los sesgos de evaluación. Sin embargo, nuestras estimaciones sugieren que la penalización crítica hacia las películas dirigidas por mujeres se reduce bajo escalas con mayor granularidad, no menor. Mientras que bajo la escala de 5 puntos los filmes dirigidos por mujeres reciben entre 0,15 y 0,19 estrellas menos en promedio, la diferencia desaparece cuando se permite el uso de medias estrellas. Esto sugiere que los críticos cinematográficos no emplean la mayor granularidad de la escala para discriminar de forma encubierta el trabajo de las directoras, como encuentran Rivera y Tilcsik (2019) y Botelho *et al*. (2025) en otros ámbitos.

Nuestro análisis tampoco confirma la hipótesis de que una escala más fina refuerza la asociación entre la máxima puntuación y atributos estereotípicamente masculinos como "brillantez" o "genialidad". Más allá del efecto mecánico de que una película tenga una mayor probabilidad de recibir la máxima nota bajo la escala de 5 puntos que bajo la de 9, no detectamos una penalización hacia las directoras en la probabilidad de recibir 5 estrellas, ni que esta dependa del tipo de escala. Una posible explicación es que, al contrario de lo que sucedía en el cuasiexperimento de Rivera y Tilcsik (2019), la puntuación máxima es de 5 estrellas en ambas escalas, por lo que el valor simbólico asociado a ese umbral no cambia.

Por último, los resultados confirman que la penalización procede principalmente de los críticos varones, en línea con lo documentado en Cubel y Sanchez-Pages (2025). Bajo escalas de 5 puntos, estos asignan entre 0,18 y 0,20 estrellas menos a las películas dirigidas por mujeres. Con medias estrellas, esa penalización desaparece. Si bien este patrón no es robusto a la introducción de efectos fijos de película y no se repite cuando estudiamos la probabilidad de recibir la máxima calificación, la evidencia es consistente en asociar el sesgo de género en las evaluaciones con los críticos hombres. Las críticas mujeres no muestran diferencias sistemáticas en ninguna especificación.

Por tanto, nuestros resultados corroboran que la arquitectura de la evaluación no es un aspecto neutral. La granularidad de la escala, lejos de ser un aspecto trivial, condiciona la magnitud y visibilidad de los sesgos de género en la crítica cinematográfica. Aportamos así evidencia de que es importante prestar atención al diseño de los sistemas de puntuación, pues modula el grado en que prejuicios y estereotipos se reflejan en la valoración del trabajo creativo.

II. REVISIÓN DE LA LITERATURA

El presente artículo está relacionado con el prolongado debate en psicometría sobre qué escala maximiza la fiabilidad y validez de las respuestas. Preston y Colman (2000) encontraron que la fiabilidad tiende a aumentar con el número de puntos, alcanzando su máximo en torno a siete categorías. Estudios posteriores matizaron esta conclusión. Cook y Beckman (2009) mostraron que una escala de 9 puntos permite distinguir mejor los niveles de competencia que una de 5, aunque ambas ofrecen una consistencia aceptable. Revilla *et al*. (2014) coincidieron en que las escalas intermedias (de siete a diez categorías) equilibran validez y precisión, mientras que Weathers *et al*. (2005) y Weijters *et al*. (2010) destacaron que las escalas muy largas generan fatiga cognitiva, por lo que pueden resultar poco informativas. Más recientemente, Yoon (2024) ha concluido que no existe un formato universalmente superior: la longitud idónea depende del contexto, de los objetivos de la medición y de las características de los encuestados. En cualquier caso, la literatura psicométrica subraya que el diseño de una escala de evaluación no es neutro.

Esa falta de neutralidad ha llevado a examinar cómo la arquitectura de la evaluación interactúa con sesgos cognitivos y sociales. Weijters *et al*. (2010) mostraron que etiquetar todas las categorías incrementa el sesgo de conformidad, pero reduce tanto el uso de extremos como las inconsistencias, mientras que la inclusión de un punto medio favorece respuestas moderadas. Eutsler y Lang (2015) observaron que las escalas completamente descritas reducen la ambigüedad, mientras las que solo tienen los extremos etiquetados amplifican la interpretación subjetiva de las posiciones intermedias.

Varios trabajos en esta literatura demuestran que la arquitectura de evaluación condiciona la expresión de prejuicios latentes. Por ejemplo, Culpepper *et al*. (2023) observaron que las rúbricas pueden mitigar sesgos en la contratación académica al obligar a los evaluadores a justificar sus puntuaciones, aunque advierten que, si los criterios son sesgados, esta herramienta puede reproducir la desigualdad. En la misma línea, Brooke y Rao (2024) demostraron que intervenciones en plataformas digitales, como simplificar escalas o destacar atributos positivos, pueden alterar sustancialmente los patrones de discriminación.

Dos trabajos se han centrado en la relación entre el diseño de las escalas de evaluación y la expresión de sesgos de género. En un cuasiexperimento natural

en una universidad estadounidense, Rivera y Tilcsik (2019) observaron que al pasar de una escala de 10 a una de 6 puntos desaparecía la brecha de género en las puntuaciones del profesorado, especialmente en áreas dominadas por hombres. Un experimento de encuesta replicó este resultado manteniendo constante la calidad de la enseñanza: la penalización a las profesoras era significativa con una escala de 10 puntos, pero desaparecía con una de 6. Los autores interpretan este fenómeno señalando que el "10" está culturalmente asociado con la brillantez y la perfección, atributos tradicionalmente más vinculados a los hombres (Leslie *et al.*, 2015), lo que hace que los evaluadores sean más reacios a asignarlo a mujeres. En cambio, el máximo en una escala del uno al seis carece de esa connotación.

Recientemente, Botelho *et al*. (2025) han analizado evaluaciones en mercados laborales digitales y encuentran que la dicotomización de escalas reduce la discriminación racial y contribuye a disminuir desigualdades de ingresos. Rivera (2025) interpreta esto como el resultado de que las escalas binarias eliminan el margen para expresar sesgos sutiles. Estos estudios sugieren, por tanto, que los sesgos se producen en los matices intermedios de las escalas, donde un evaluador puede justificar otorgar un "4" en lugar de un "5" sin reconocerlo como un acto discriminatorio.

Este marco es pertinente para la crítica cultural y artística, especialmente la cinematográfica, donde abundan las puntuaciones numéricas. La historiografía del cine ha consagrado a directores varones como paradigmas del genio creador, mientras que las mujeres directoras han sido invisibilizadas o tratadas como excepciones (Nochlin, 1972). Esta pauta se ha mantenido gracias a un ecosistema cinematográfico en el que las mujeres continúan infrarrepresentadas tanto en puestos creativos clave como en la crítica (Loist *et al.*, 2024; VIDA Foundation, 2020).

La investigación reciente confirma que los sesgos de género afectan la valoración de obras creativas en múltiples campos, incluido el cine. Cubel y Sanchez-Pages (2025) muestran que las películas dirigidas por mujeres reciben calificaciones más bajas en promedio, y que esta penalización proviene sobre todo de los críticos varones. Su análisis textual de reseñas críticas revela que los filmes de directoras tienen menos probabilidades de ser descritos como obras "brillantes" o "excepcionales" y que se les asocia con más frecuencia a temas subjetivos y particulares.

Patrones comparables emergen en la crítica literaria. Los críticos suelen favorecer a autores de su mismo género, y se ha observado que las críticas profesionales valoran relativamente peor a autoras. Además, las obras de mujeres se asocian con géneros y temas de menor prestigio, mientras que las de hombres se evalúan en términos de estructura y estilo (Lassen *et al.*, 2022, 2023; Thelwall, 2019; Touileb *et al.*, 2020). Estas diferencias reproducen el estereotipo cultural que identifica la creatividad con lo masculino (Proudfoot *et al.*, 2015). No sorprende, por ello, que intervenciones como las audiciones "a ciegas" en orquestas sirvan para aumentar la contratación de mujeres (Goldin y Rouse, 2000).

Este artículo estudia la intersección entre crítica cultural y arquitectura de la evaluación. Nuestro objetivo es aplicar este marco teórico y empírico al campo del cine. A diferencia de otros ámbitos como la docencia universitaria o las plataformas de trabajo temporal, en los que los evaluadores son un grupo mucho más numeroso que los evaluados, nuestro análisis se centra en un colectivo muy definido de evaluadores profesionales: los críticos.

III. DATOS

Recopilamos una muestra de 4.201 reseñas de largometrajes publicadas entre agosto de 2021 y agosto de 2025 en cuatro medios digitales españoles. Dos de ellos utilizan calificaciones que solo permiten estrellas enteras, de 1 a 5: *Fotogramas.es* y *El Antepenúltimo Mohicano* (*EAM*). *Fotogramas* es la revista cinematográfica más antigua de España, fundada en 1946. Su web recibe más de seis millones de usuarios únicos mensuales (ComScore, 2024). Todas sus reseñas están firmadas por críticos profesionales. EAM, por su parte, es un portal independiente dedicado al cine de autor y de festivales. Aunque sus colaboradores no siempre cuentan con una carrera periodística establecida, sus textos presentan un marcado tono analítico y académico, como confirman los indicadores lin-

güísticos de legibilidad y complejidad léxica que se mostrarán más adelante. EAM opera a menor escala, se estima que recibe entre 80.000 y 150.000 visitas únicas mensuales, pero goza de prestigio en la cultura cinéfila en castellano por su atención constante al cine independiente y de autor.

Los otros dos medios, *Cinemanía* y *La Razón*, utilizan escalas que permiten medias estrellas y que, por tanto, tienen 9 puntos posibles. *Cinemanía* fue fundada en 1995 y tiene una presencia consolidada como revista cinematográfica especializada, con ediciones impresas y digitales. Es segunda tras *Fotogramas* en terminos de usuarios únicos dentro de la categoría de crítica de cine (ComScore, 2024). Sus reseñas son elaboradas por críticos de plantilla y colaboradores regulares, y cubren tanto estrenos comerciales como títulos de cine de autor. *La Razón*, en cambio, es un diario generalista con una sección cultural y de espectáculos que incluye críticas cinematográficas. Sus reseñas, aunque menos extensas, aportan una perspectiva distinta: su tono es más periodístico, adaptado a un público amplio, y permiten evaluar si el sesgo de género detectado en espacios especializados se reproduce también en medios generalistas.

En conjunto, la muestra incluye reseñas de 1.919 películas distintas, escritas por 105 críticos, de los cuales el 27,6 por 100 son mujeres. Para cada película, hemos recopilado variables en dos bloques principales: características de producción y distribución, e indicadores de recepción. Las primeras incluyen género, duración, países de producción, lenguas y tipo de estreno, obtenidos de la base de datos abierta The Movie Database (TMDb). También medimos la participación de mujeres en guion y dirección de fotografía. Cuando la información de género del equipo creativo no estaba disponible en TMDb, la completamos manualmente revisando perfiles profesionales, entrevistas y páginas oficiales.

Los indicadores de recepción provienen de la popular web IMDb, e incluyen el número de votos y calificación media como aproximación al reconocimiento del público, y el recuento de premios y nominaciones como medida de validación artística. También recopilamos datos de taquilla mundial y presupuesto de producción. Los datos de ingresos proceden de IMDb y, cuando fue necesario, de The Numbers, TMDb y el Instituto de la Cinematografía y de las Artes Audiovisuales (ICAA). Cabe mencionar que la información sobre presupuestos de producción presenta vacíos importantes, especialmente concentrados en producciones independientes.

Por último, para caracterizar el estilo de las reseñas, aplicamos dos métricas de legibilidad en castellano: el índice de perspicuidad de Szigriszt-Pazos (1993), basado en la longitud de frases y el cómputo de sílabas, y el índice μ de legibilidad (Muñoz, 2006), que mide la complejidad léxica a partir de la longitud y variación de las palabras.

IV. ANÁLISIS DESCRIPTIVO

1. Por escala

Comenzamos examinando las diferencias de calificaciones entre escalas de estrellas enteras y aquellos que incorporan medias estrellas (cuadro n.º 1). Las reseñas en una escala de 5 puntos otorgan una media de 3,45 estrellas, significativamente superior a la de los medios que permiten medias estrellas, 3,33. En la primera escala, el 7,8 por 100 de las reseñas asigna la máxima calificación de 5 estrellas, mientras que en la escala de 9 puntos la proporción se reduce al 4 por 100. Esta diferencia es esperable dado el efecto mecánico de contar con un mayor número de categorías, que hace menos frecuente el uso de la máxima calificación.

En cuanto al perfil de género, no aparecen diferencias significativas entre los dos grupos de medios. Las películas dirigidas por mujeres representan el 24,4 por 100 de las reseñadas en medios que utilizan estrellas enteras y el 22,6 por 100 en los de medias estrellas. Lo mismo ocurre con la proporción de mujeres en el equipo de guionistas (29 por 100 vs. 28,2 por 100) y de directoras de fotografía (13,8 por 100 vs. 13,3 por 100). Donde sí se observa una diferencia significativa es en la autoría crítica: las mujeres firman el 17,5 por 100 de las reseñas en los medios con estrellas enteras, frente al 29,9 por 100 en los que permiten medias estrellas.

Un aspecto adicional es la consistencia de las valoraciones entre críticos que reseñan una misma pe-

lícula, que puede medirse estimando la correlación intraclase (ICC). En la escala de estrellas enteras, la ICC alcanza un valor de 0,43, lo que implica que cerca del 43 por 100 de la variación en las puntuaciones se debe a diferencias entre películas, mientras que el resto obedece a diferencias sistemáticas entre autores y a variación idiosincrática. En cambio, en las cabeceras que utilizan medias estrellas, la ICC desciende a 0,24, reflejando un menor grado de acuerdo entre críticos (1).

CUADRO N.º 1

ESTADÍSTICOS DESCRIPTIVOS DE RESEÑAS POR TIPO DE ESCALA

VARIABLE	ESTRELLAS ENTERAS (*n* = 2027)	MEDIAS ESTRELLAS (*n* = 2174)
Estrellas	3,45 [3,42, 3,49]	3,33 [3,30, 3,36]
Legibilidad	61,78 [61,54, 62,02]	51,62 [51,35, 51,89]
Simplicidad	56,24 [56,06, 56,41]	37,07 [36,73, 37,40]
Críticas mujeres	17,5% [15,9%, 19,2%]	29,9% [28,0%, 31,9%]
Directoras mujeres	24,7% [22,9%, 26,6%]	23,0% [21,2%, 24,8%]
Guionistas mujeres	29,0% [27,0%, 31,0%]	28,2% [29,0%, 34,8%]
Directoras de fotografía	13,8% [12,3%, 15,3%]	13,3% [11,9%, 14,8%]
Nota IMDb	6,50 [6,47, 6,53]	6,40 [6,36, 6,43]
Votos en IMDb	43.096 [38.597, 56.972]	39.105 [34.979, 43.231]
Taquilla	$89,35M [$75,26M, $103,45M]	$92,86M [$78,37M, $107,35M]]
Presupuesto	$63,27M [$53,44M, $73,11M]	$47,23M [$42,61M, $51,85M]
Premios	10,94 [9,69, 12,18]	9,31 [8,26, 10,36]
Nominaciones	26,45 [24,05, 28,85]	23,03 [20,83, 25,24]
Drama	67,4% [59,3%, 65,3%]	61,8% [69,5%, 75,1%]
Comedia	24,1% [22,3%, 26,0%]	27,1% [25,3%, 29,0%]
Producción EE. UU.	37,5% [35,4%, 39,6%]	33,9% [31,9%, 35,9%]
Producción España	21,9% [20,1%, 23,7%]	27,9% [26,1%, 29,8%]

La desagregación por cabecera (cuadro n.º A1) muestra diferencias que reflejan culturas críticas diferenciadas. *Fotogramas* es el medio más generoso en sus puntuaciones (3,53 estrellas de media), y con un estilo de escritura más accesible (legibilidad de 63,8, simplicidad de 57,3). *EAM* se distingue por un lenguaje más denso (59,7 de legibilidad, 55,2 de simplicidad) y por reseñar más cine de autor, con la mayor proporción de directoras (27,9 por 100), guionistas mujeres (31,8 por 100) y directoras de fotografía (15,5 por 100). *Cinemanía* se sitúa en una posición intermedia (3,40 estrellas otorgadas en media), pero tiene el estilo más complejo (legibilidad de 52,0), y un mayor peso de las producciones nacionales (30,6 por 100). *La Razón* es el medio más estricto en sus puntuaciones (3,19) pese a ser también la cabecera que cubre filmes de mayor prestigio, medido por el número de nominaciones y premios (28,7 nominaciones y 11,6 premios de media). Aunque *La Razón* destaca por el alto peso de su crítica femenina (32,4 por 100 de las reseñas), hay que mencionar que su sección de crítica cinematográfica está formada únicamente por dos periodistas, un hombre y una mujer.

2. Por género del director

De las 1.919 películas reseñadas, el 23,8 por 100 fueron dirigidas por mujeres, el 73,3 por 100 por hombres y el resto por equipos mixtos. Estos porcentajes se mantienen al considerar el universo de críticas (23,8 por 100 dirigidas por mujeres, 76 por 100 por hombres). El cuadro n.º 2 muestra que, agregando ambas escalas, las películas dirigidas por mujeres reciben valoraciones significativamente más bajas que las de hombres (3,34 frente a 3,40, $z = 2,08$, $p = 0,038$) y reciben una proporción menor de máximas calificaciones (4,3 por 100 frente a 6,3 por 100; $z = 2,38$, $p = 0,017$).

El gráfico 1 muestra la distribución de las calificaciones en estrellas por género del director bajo cada escala. En el caso de estrellas completas

CUADRO N.º 2

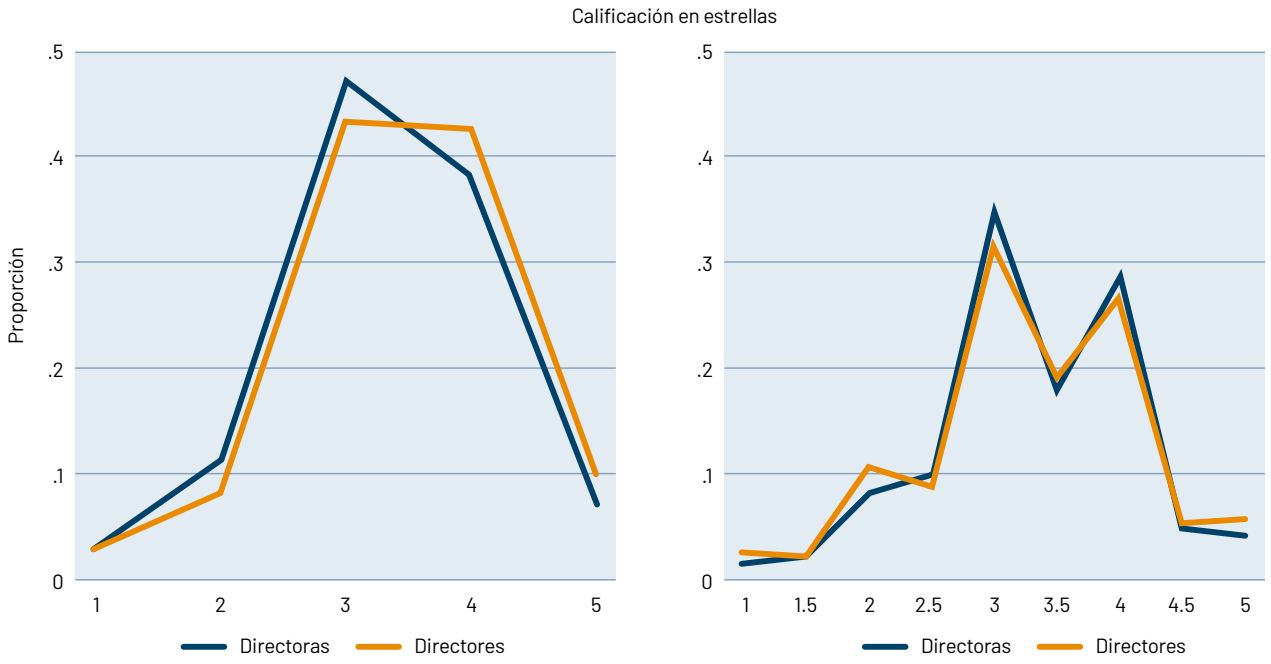
ESTADÍSTICAS DESCRIPTIVAS POR GÉNERO DE DIRECTOR Y CRÍTICO

VARIABLE	DIRIGIDAS POR HOMBRES (*n* = 3096)	DIRGIDADAS POR MUJERES (*n* = 1000)	CRÍTICOS HOMBRES (*n* = 3189)	CRÍTICAS MUJERES (*n* = 1004)
Estrellas	3,40 [3,37, 3,42]	3,35 [3,30, 3,39]	3,37 [3,34, 3,40]	3,45 [3,41, 3,49]
5 estrellas	6,33 [5,47,7,19]	4,30 [3,04,5,56]	5,96 [5,14,6,78]	5,38 [3,98,6,77]
Legibilidad	56,69 [56,41, 56,96]	56,08 [55,58, 56,58]	56,38 [56,10, 56,65]	56,99 [56,53, 57,46]
Simplicidad	46,40 [45,99, 46,80]	46,09 [45,37, 46,81]	46,85 [46,46, 47,25]	44,60 [43,87, 45,33]
Guionistas mujeres	11,1% [10,1%, 12,3%]	80,5% [77,8%, 82,7%]	26,8% [25,2%, 28,3%]	34,2% [31,3%, 37,2%]
Directoras de fotografía	8,6% [7,6%, 9,7%]	28,5% [25,8%, 31,4%]	12,9% [11,7, 14,1%]	15,7% [13,5%, 18,1%]
Nota IMDb	6,46 [6,43, 6,49]	6,40 [6,35, 6,45]	6,45 [6,42, 6,48]	6,44 [6,39, 6,49]
Votos en IMDb	49.048 [45.160, 52.936]	18.894 [15.010, 22.778]	44.175 [40.478, 47.873]	31.064 [26.186, 35.942]
Taquilla	$103,19M [$91,22M, $115,15M]	$35,29M [$20,25M, $50,34M]	$96,28M [$84,32M, $108,25M]	$72,35M [$54,56M, $90,14M]
Presupuesto	$52,15M [$48,39M, $55,91M]	$24,41M [$18,35M, $30,48M]	$49,30M [$45,53M, $53,07M]	$41,18M [$34,50M, $47,87M]
Premios	10,22 [9,27, 11,17]	10,22 [8,54, 11,90]	10,18 [9,22, 11,14]	9,87 [8,40, 11,34]
Nominaciones	26,07 [24,08, 28,06]	21,62 [18,68, 24,56]	24,77 [22,89, 26,66]	24,50 [21,25, 27,75]
Drama	60,8% [59,1%, 62,5%]	76,3% [73,6%, 78,8%]	62,7% [61,1%, 64,4%]	70,1% [67,2%, 72,9%]
Comedia	27,0% [25,4%, 28,6%]	20,5% [18,1%, 23,1%]	25,5% [24,0%, 27,1%]	26,1% [23,5%, 28,9%]
Producción EE. UU.	39,3% [37,6%, 41,0%]	24,8% [22,2%, 27,6%]	37,6% [35,9%, 39,3%]	29,4% [26,6%, 32,3%]
Producción España	22,9% [21,4%, 24,4%]	32,2% [29,4%, 35,2%]	23,2% [21,8%, 24,7%]	30,6% [27,8%, 33,5%]

(panel izquierdo), el desplazamiento hacia notas más bajas para las directoras es evidente. La escala de medias estrellas (panel derecho) presenta una distribución más compleja. La comparación entre géneros en este caso sugiere una mayor varianza en las calificaciones de las películas dirigidas por hombres, que reciben con mayor frecuencia notas extremas (1-1,5 y 5 estrellas). Esta diferencia en variabilidad es estadísticamente significativa (test de Levene, $p = 0{,}019$). Las películas reseñadas dirigidas por mujeres reciben también menores puntuaciones en IMDb (6,39 frente a 6,46; $t = 2{,}29$, $p = 0{,}022$) y son menos populares (19.027 frente a 47.778 votos recibidos en IMDb; $t = 7{,}91$, $p < 0{,}001$). También cuentan con presupuestos y taquillas significativamente inferiores ($t = 1{,}96$, $p = 0{,}050$ para presupuestos,

GRÁFICO 1

DISTRIBUCIÓN DE VALORACIONES POR TIPO DE ESCALA



Fuente: Elaboración propia.

t = 6,43, p < 0,001 para ingresos). En términos de reconocimiento artístico, las películas dirigidas por mujeres reciben menos nominaciones (t = 2,16, p = 0,031), pero no menos premios (t = -0,09, p = 0,926). Las mujeres directoras dirigen más dramas (76,3 por 100 frente a 60,9 por 100) y menos comedias (20,8 por 100 frente a 26,8 por 100).

Otra diferencia marcada aparece en la composición de género de los equipos creativos: en las películas dirigidas por mujeres, el 80,5 por 100 de los guionistas y el 28,1 por 100 de los directores de fotografía son mujeres, frente al 11,5 por 100 y 8,8 por 100, respectivamente, en las películas dirigidas por hombres. Estos patrones reflejan una mayor propensión de las directoras a colaborar con otras mujeres que también aparece en otros entornos (Ductor *et al*., 2023).

3. Por género del crítico

Los críticos mujeres firman el 24 por 100 de las reseñas. Se observa que, bajo la escala de estrellas enteras, los críticos hombres son significativamente más estrictos en sus valoraciones que sus colegas mujeres (3,37 frente a 3,45 estrellas; z = 4,57, p < 0,001), mientras que en la escala con medias estrellas no hay diferencia (z = 0,49, p = 0,625). En relación con la máxima calificación, en la escala de estrellas enteras se mantiene que los hombres son más estrictos (7,3 por 100 frente a 10,2 por 100), aunque la diferencia es marginalmente significativa (z = 1,83, p = 0,067). Bajo la escala de 9 puntos, ocurre lo contrario: los hombres son más propensos a otorgar la máxima puntuación (4,5 por 100 frente a 2,8 por 100), y la significación es más clara (z = -1,87, p = 0,061).

En cuanto a las películas reseñadas, los críticos hombres cubren filmes de similar presupuesto y recepción critica que los reseñados por mujeres (2), pero que son más populares (44.175 frente a 31.120 votos en IMDb; t= -3,60, p < 0,001) y recaudan significativamente más (54,6 frente a 37 millones de dólares; t = -2,60, p = 0,009). Por el contrario,

las críticas mujeres reseñan películas con mayor presencia femenina en sus equipos creativos que los críticos hombres (34,2 por 100 de guionistas frente a 26,8 por 100; 15,7 por 100 de directoras de fotografía frente a 12,9 por 100). También cubren más dramas (70,1 por 100 frente a 62,7 por 100), más cine español (30,6 por 100 frente a 23,2 por 100) y menos cine estadounidense (29,4 por 100 frente a 37,6 por 100).

Estas diferencias reflejan la fuerte correlación de género en la asignación editorial de películas. Las críticas mujeres reseñan en mayor medida películas dirigidas por mujeres que sus colegas hombres (29,7 por 100 frente a 22 por 100; $\chi^2(1) = 24{,}7$, $p < 0{,}001$). Esta diferencia se reproduce tanto en los medios con estrellas enteras (33,5 por 100 frente a 23,6 por 100; $\chi^2(1) = 14{,}8$, $p < 0{,}001$) como en los que emplean medias estrellas (28,9 por 100 frente a 21,2 por 100; $\chi^2(1) = 14{,}6$, $p < 0{,}001$).

También se observan diferencias de género en el estilo de escritura: las reseñas firmadas por mujeres son más legibles (57 frente a 56,4; $t = 2{,}14$, $p = 0{,}032$), pero menos simples léxicamente (44,6 frente a 46,9; $t = -5{,}33$, $p < 0{,}001$).

En conjunto, estos patrones apuntan a que las diferencias en las valoraciones otorgadas por críticos hombres y mujeres no pueden atribuirse a una disparidad en calidad objetiva, sino, en todo caso, a una división de trabajo por género en la crítica cinematográfica. Las críticas mujeres tienden a cubrir filmes menos comerciales y con equipos más inclusivos, mientras que los críticos hombres reseñan producciones de mayor popularidad y mejores taquillas. Esta división se refleja también en la asignación según el género del equipo directoral: las críticas mujeres reseñan proporcionalmente más películas dirigidas por mujeres, una diferencia que se reproduce bajo ambos sistemas de puntuación.

V. RESULTADOS

1. Análisis univariante

Comenzamos con un conjunto de pruebas no paramétricas para examinar si la arquitectura de la escala de puntuación modera las diferencias de valoración entre películas dirigidas por mujeres y por hombres. Dado que las puntuaciones en estrellas son categorías discretas y no se distribuyen normalmente, usamos pruebas de Mann-Whitney para comparar distribuciones entre grupos.

En el conjunto de reseñas que emplean únicamente estrellas enteras, las películas dirigidas por mujeres reciben una puntuación significativamente más baja que las dirigidas por hombres ($z = -3{,}07$, $p = 0{,}0021$). Esta diferencia desaparece en las reseñas que utilizan escalas con medias estrellas ($z = 0{,}19$, $p = 0{,}853$). Esta comparación inicial sugiere que la penalización hacia el cine dirigido por mujeres deja de ser visible cuando los críticos disponen de mayor granularidad para matizar sus juicios.

Al desagregar por género del crítico, se observa que la diferencia en la escala de estrellas enteras se debe a que los críticos varones otorgan menores puntuaciones a las películas dirigidas por mujeres ($z = -3{,}70$, $p < 0{,}001$). En las críticas firmadas por mujeres, en cambio, no hay diferencias ($z = -0{,}56$, $p = 0{,}573$). En la escala con medias estrellas, ni críticos hombres ($z = -0{,}75$, $p = 0{,}452$) ni críticas mujeres ($z = 1{,}50$, $p = 0{,}134$) muestran diferencias significativas en sus valoraciones según el género del equipo directoral.

Otra medida de interés es la proporción de películas consideradas brillantes o magistrales y, por tanto, merecedoras de recibir la puntuación máxima de 5 estrellas. En la escala de estrellas enteras, las películas dirigidas por mujeres alcanzan esta calificación con menor frecuencia que las dirigidas por hombres (5,8 por 100 frente a 8,5 por 100), una diferencia cercana a la significación estadística ($z = -1{,}94$, $p = 0.053$). En la escala con medias estrellas, la brecha deja de ser significativa ($z = -1{,}57$, $p = 0{,}117$). Al considerar el género del director, las pruebas univariantes sugieren que la menor frecuencia de 5 estrellas para las películas dirigidas por mujeres de nuevo proviene de los críticos varones. En la escala de medias estrellas, solo un 2,5 por 100 de estas películas recibe la puntuación máxima, frente al 5,1 por 100 de las películas de directores hombres ($z = -1{,}95$, $p = 0{,}051$). La diferencia es menor y se sitúa en el margen de significación bajo la escala de estrellas enteras ($z = -1{,}59$,

p = 0,111). Entre las críticas escritas por mujeres no hay evidencia de diferencias significativas bajo ninguno de los dos sistemas.

Estos análisis sugieren que el sesgo contra películas dirigidas por mujeres está presente sobre todo cuando las películas se califican bajo una escala de estrellas enteras, y se concentra en los críticos hombres. La escala de medias estrellas parece atenuar ese patrón, permitiendo que las valoraciones de filmes dirigidos por mujeres converjan con las de sus pares masculinos. Como en el análisis previo, cabe recordar que las diferencias observadas podrían reflejar variaciones sistemáticas en el tipo de filmes reseñados. Para discernir si se trata de un efecto de arquitectura de escala o de selección de películas, pasamos a continuación al análisis multivariante.

2. Análisis de regresión

Calificaciones

Comenzamos estimando modelos de mínimos cuadrados ordinarios (MCO) con efectos aleatorios y errores estándar a nivel de crítico en los que la variable dependiente es la calificación en estrellas otorgada a cada película. Nuestro interés recae en dos variables: una dicotómica que, identifica si la película ha sido dirigida por una mujer, y un indicador de la escala de puntuación, que distingue entre el sistema con solo estrellas completas (escala de 5 puntos), y el que permite medias estrellas (9 puntos). El coeficiente clave es el del término de interacción entre estas dos variables, pues nos permite responder a la pregunta de hasta qué punto la penalización sistemática contra las directoras observada en Cubel y Sanchez-Pages (2025) se modula en función del tipo de escala utilizada. El cuadro n.º 3 muestra los resultados.

El modelo en la columna [1] incluye todas las reseñas para las que contamos con un conjunto amplio de controles, tanto de características de la película como de su recepción. Este modelo indica la existencia de un sesgo negativo contra las directoras bajo la escala de estrellas completas (-0,124, p = 0,011) y la atenuación de ese sesgo bajo la escala de medias estrellas (0,085, p = 0,108).

La evidencia se refuerza cuando añadimos efectos fijos de cabecera (columna 2). El coeficiente asociado a la variable dicotómica de dirección femenina indica que, a igualdad de condiciones, las películas dirigidas por mujeres reciben, en promedio, 0,15 estrellas menos (un 4 por 100 de la media) que las dirigidas por hombres bajo la escala de estrellas enteras. El término de interacción con la escala con medias estrellas es positivo y significativo (0,115, p = 0,026). Esto significa que la penalización hacia las mujeres se reduce en presencia de una escala más fina. De hecho, el efecto marginal de que la directora sea mujer en una escala de 9 puntos se reduce a -0,039 y deja de ser significativo. En otras palabras, la posibilidad de otorgar medias estrellas neutraliza la penalización crítica hacia los filmes dirigidos por mujeres.

En el resto de las columnas restringimos la muestra a las 1.082 películas que fueron reseñadas al menos una vez bajo cada tipo de escala, reduciendo así el posible sesgo en la selección de películas. En la columna [3], con controles y efectos fijos de cabecera, la penalización a las directoras aumenta (-0,184, p < 0,01). La interacción con la escala de medias estrellas es positiva, pero menos precisa (0,106, p = 0,074). El patrón de atenuación persiste, aunque la diferencia entre sistemas es menos marcada que en las especificaciones anteriores. De hecho, el efecto marginal de que la película esté dirigida por una mujer en sistemas de 9 puntos es ahora similar en magnitud y significatividad al estimado para la escala de 5 puntos en los anteriores modelos (-0,086, p = 0,009).

Las columnas [4] y [5] introducen efectos fijos de película, que comparan cómo los críticos puntúan exactamente el mismo filme bajo cada escala. Dado que en este caso el efecto de la dirección femenina se absorbe por los efectos fijos, estimamos regresiones separadas por género del director. En la submuestra de películas dirigidas por mujeres (columna 4), el coeficiente de la escala de 9 puntos es cercano a 0 e insignificante (0,006; p = 0,918). En cambio, en la submuestra de películas dirigidas por hombres (columna 5), la escala de 9 puntos reduce las calificaciones en -0,076 (p = 0,081). Esto sugiere que la menor diferencia en puntuación que observamos bajo la escala más fina no se debe a una mejora en las notas recibidas por las mujeres, sino a un descenso en las valoraciones otorgadas a los hombres. Sin embargo, este resultado

CUADRO N.º 3

REGRESIONES DE CALIFICACIÓN POR GÉNERO DEL DIRECTOR

VARIABLE	[1]	[2]	[3]	[4]	[5]
Director mujer	-0,124**	-0,154***	-0,184***	—	—
	(0,048)	(0,050)	(0,055)	—	—
Medias estrellas	-0,229*	-0,137***	-0,073	0,000	-0,076*
	(0,119)	(0,053)	(0,052)	(0,059)	(0,043)
Director mujer x medias estrellas	0,086	0,115**	0,098*	—	—
	(0,053)	(0,057)	(0,059)	—	—
Crítico hombre	-0,155**	-0,086**	-0,070*	-0,104*	-0,079
	(0,069)	(0,042)	(0,041)	(0,055)	(0,057)
Nominaciones	0,003***	0,003***	0,003***	—	—
	(0,000)	(0,000)	(0,000)	—	—
Premios	-0,001*	-0,002**	-0,001	—	—
	(0,001)	(0,001)	(0,001)	—	—
Log(votos IMDb)	0,075***	0,087***	0,086***	—	—
	(0,015)	(0,016)	(0,018)	—	—
Nota IMDb	0,203***	0,214***	0,189***	—	—
	(0,023)	(0,025)	(0,029)	—	—
Log(taquilla)	-0,036***	-0,036***	-0,048***	—	—
	(0,009)	(0,010)	(0,011)	—	—
EF cabecera	NO	SÍ	SÍ	SÍ	SÍ
EF película	NO	NO	NO	SÍ	SÍ
Observaciones	3.904	3.904	3.188	775	2.403

Notas: Regresiones MCO. Errores agrupados a nivel de crítico. Los modelos [1] a [3] incluyen variables de duración del filme, tipo de estreno, género, idioma y países de producción. * $p < 0,10$, ** $p < 005$, *** $p < 0,01$.

debe interpretarse con cautela en ausencia de evidencia adicional.

En conjunto, estos resultados muestran que las escalas con más puntos reducen la penalización hacia el trabajo de mujeres directoras que se observa bajo el sistema de estrellas enteras. Bajo la escala de 5 puntos, las películas dirigidas por mujeres reciben entre 0,15 y 0,18 estrellas menos en promedio, controlando por múltiples factores. Cuando la escala permite medias estrellas, la penalización se reduce a valores en torno a 0,03 a 0,08 estrellas, y en la mayoría de las especificaciones no es estadísticamente significativo.

Resultado 1: Las películas dirigidas por mujeres reciben menores valoraciones bajo la escala de estrellas completas. Esta penalización se atenúa e incluso desaparece cuando la escala permite medias estrellas.

La probabilidad de obtener cinco estrellas

A continuación, estimamos una serie de modelos lineales de probabilidad (MLP) en los que la variable dependiente es un indicador de si la reseña asigna la máxima puntuación de 5 estrellas. El interés de esta especificación es doble. Por un lado, muestra los determinantes del umbral cualitativo que los críticos emplean para considerar un trabajo como excelente o magistral, atributos que suelen asociarse con lo masculino (Nochlin, 1971; Leslie *et al.*, 2015). En segundo lugar, nos permiten examinar si, como sugieren Rivera y Tilcsik (2019), el sesgo de género en el reconocimiento de la excelencia es mayor cuando la escala es más fina.

Los resultados se muestran en el cuadro n.º 4. Una primera observación es que el uso de escalas de 9 puntos reduce la probabilidad de recibir la máxima calificación. En todos los modelos, el coeficiente de

CUADRO N.º 4

PROBABILIDAD DE OBTENER 5 ESTRELLAS POR GENERO DEL DIRECTOR

VARIABLE	[1]	[2]	[3]	[4]	[5]
Director mujer	-0,025 (0,016)	-0,026 (0,016)	-0,026 (0,017)	— —	— —
Medias estrellas	-0,033** (0,013)	-0,029* (0,016)	-0,027 (0,017)	-0,025 (0,016)	-0,076* (0,043)
Director mujer x medias estrellas	0,012 (0,018)	0,013 (0,018)	0,010 (0,020)	0,010 (0,016)	— —
Crítico hombre	-0,001 (0,014)	-0,001 (0,014)	0,003 (0,015)	0,003 (0,014)	-0,079 (0,057)
Nominaciones	0,001*** (0,000)	0,001*** (0,000)	0,001*** (0,000)	— —	— —
Premios	-0,000 (0,000)	-0,000 (0,000)	-0,001 (0,000)	— —	— —
Log(votos IMDb)	0,015*** (0,003)	0,015*** (0,003)	0,016*** (0,004)	— —	— —
Nota IMDb	0,008 (0,005)	0,008 (0,005)	0,011* (0,007)	— —	— —
Log(taquilla)	-0,003 (0,002)	-0,003 (0,002)	-0,003 (0,003)	— —	— —
EF cabecera	NO	SÍ	SÍ	SÍ	SÍ
EF película	NO	NO	NO	SÍ	SÍ
Observaciones	3.904	3.904	3.188	3.247	2.403

Notas: Regresiones MCO. Errores agrupados a nivel de crítico. Los modelos [1] a [3] incluyen variables de duración del filme, tipo de estreno, género, idioma y países de producción. * $p < 0,10$, ** $p < 005$, *** $p < 0,01$.

la variable de escala es negativo y en varios casos alcanza significación estadística. Este patrón es esperable ya que, cuanto menor es el número de categorías disponibles, mayor es la probabilidad de que se asigne la máxima nota.

Nuestro interés principal recae en la interacción entre género del equipo directoral y tipo de escala. Aquí, la evidencia es más débil. En el modelo con controles y la muestra completa (columna 1), el efecto marginal de que la directora sea mujer bajo la escala de 5 puntos es negativo, pero no alcanza niveles convencionales de significación (-0,025, $p = 0,122$). La interacción con la escala de 9 puntos sugiere cierta atenuación, pero no es significativa (0,012, $p = 0,512$). El efecto marginal bajo la escala de 9 puntos es de -0.,013 y solo marginalmente significativo ($p = 0,158$).

La inclusión de efectos fijos de cabecera (columna 2) no altera este resultado: el coeficiente principal se mantiene negativo (-0,026, $p = 0,108$), y la interacción sigue siendo pequeña y no significativa (0,013, $p = 0,478$). En la muestra restringida a películas reseñadas bajo ambos sistemas (columna 3), los coeficientes apenas varían.

Finalmente, en la columna [4] introducimos efectos fijos de película. En este caso no separamos por género del director, ya que el efecto mecánico del número de puntos de la escala complica la comparación directa del sesgo de género. El coeficiente de interacción continúa sin mostrar una dirección clara (0,010; $p = 0,540$). En todas las especificaciones, el efecto marginal de la dirección femenina sobre la probabilidad de recibir 5 estrellas es pequeño, negativo y estadísticamente indistinguible de 0.

Resultado 2: No se observa una penalización de género sobre la probabilidad de que una película reciba la máxima puntuación bajo ningún tipo de escala.

CUADRO N.º 5

REGRESIONES DE CALIFICACIÓN POR GENERO DEL DIRECTOR Y DEL CRÍTICO

VARIABLE	[1]	[2]	[3]	[4]	[5]
Director mujer	-0,076 (0,097)	-0,075 (0,086)	-0,104 (0,081)	— —	— —
Medias estrellas	-0,176 (0,116)	-0,118 (0,087)	-0,026 (0,086)	0,102 (0,086)	0,033 (0,090)
Director mujer x medias estrellas	0,097 (0,108)	0,053 (0,094)	0,036 (0,086)	— —	— —
Crítico hombre	-0,092 (0,075)	-0,057 (0,065)	-0,018 (0,069)	-0,013 (0,078)	0,005 (0,062)
Crítico hombre x medias estrellas	0,047 (0,151)	-0,022 (0,092)	-0,060 (0,094)	-0,166* (0,099)	-0,143 (0,096)
Director mujer x medias estrellas x crítico hombre	0,038 (0,128)	0,077 (0,115)	0,076 (0,111)	— —	— —
Nominaciones	0,003*** (0,000)	0,003*** (0,000)	0,003*** (0,000)	— —	— —
Premios	-0,001* (0,001)	-0,002** (0,001)	-0,001 (0,001)	— —	— —
Log(votos IMDb)	0,085*** (0,015)	0,087*** (0,016)	0,087*** (0,018)	— —	— —
Nota IMDb	0,206*** (0,025)	0,214*** (0,025)	0,189*** (0,029)	— —	— —
Log(taquilla)	-0,031*** (0,010)	-0,036*** (0,010)	-0,048*** (0,011)	— —	— —
EF cabecera	NO	SÍ	SÍ	SÍ	SÍ
EF película	NO	NO	NO	SÍ	SÍ
Observaciones	3.904	3.904	3.188	775	2.403

Notas: Regresiones MCO. Errores agrupados a nivel de crítico. Los modelos [1] a [3] incluyen variables de duración del filme, tipo de estreno, género, idioma y países de producción. * $p < 0,10$, ** $p < 005$, *** $p < 0,01$.

Este resultado contrasta con el sesgo detectado en las calificaciones medias. Una interpretación plausible es que, bajo la escala de 5 puntos, un crítico que considera que la película de una directora no merece la máxima nota puede otorgarle 4 estrellas como mucho. En sistemas con medias estrellas, en cambio, dispone de una opción intermedia [4,5], lo que aumenta la calificación media sin que su propensión a reconocer una película como obra maestra difiera según el género de su director (3).

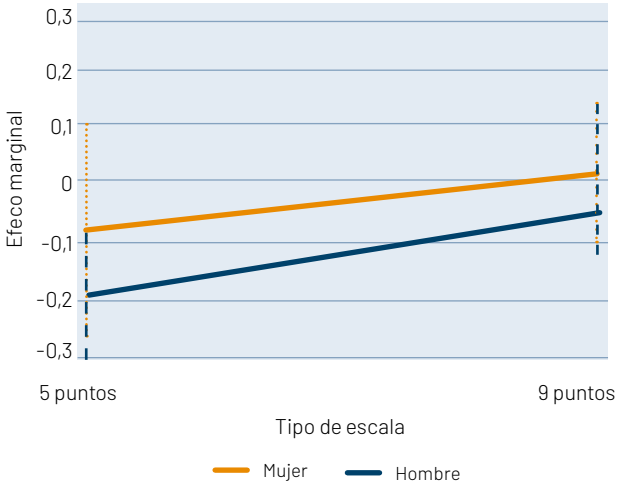
El género del crítico

Incorporamos ahora el género del crítico con el fin de examinar si la penalización observada hacia las películas dirigidas por mujeres depende del género de quién evalúa, como observaron Cubel y Sanchez-Pages (2025), y si la arquitectura de la escala modula esa penalización. Para ello, estimamos de nuevo modelos de MCO que ahora incluyen la interacción triple entre dirección femenina, tipo de escala (9 vs. 5 puntos) y si el autor de la reseña es un hombre. El cuadro n.º 5 resume los resultados.

En la especificación base con controles y muestra completa (columna 1), los coeficientes de interacción no son significativos. Los efectos marginales, que resumimos de forma visual en el gráfico 2, ofrecen una lectura más clara: bajo escalas de estrellas completas, los críticos hombres asignan en promedio unas 0,20 estrellas menos a las películas dirigidas por mujeres (p = 0,001), mientras que la diferencia es pequeña y no significativa para las críticas mu-

GRÁFICO 2

EFECTO SOBRE LA CALIFICACIÓN DE QUE LA DIRECTORA SEA MUJER POR GÉNERO DEL CRÍTICO



Fuente: Elaboración propia.

jeres, (-0,076; p = 0,434). Con medias estrellas, ningún efecto es significativo a niveles convencionales (p = 0,082 y p = 0,726, respectivamente). En suma, la penalización de género se concentra entre los críticos hombres y bajo la escala más rígida.

La inclusión de efectos fijos de cabecera y la restricción a películas reseñadas bajo ambas escalas refuerzan este patrón: la penalización por parte de los críticos hombres oscila entre -0,178 (p = 0,002) y -0,208 (p < 0,01), sin cambios para las críticas mujeres. Sin embargo, en el modelo [3], los efectos marginales muestran que la penalización por parte de los críticos hombres bajo medias estrellas, aunque atenuada, es significativa (-0,097, p = 0,011).

Al estimar modelos con efectos fijos de película (columnas 4 y 5), el panorama cambia. Para sortear la absorción del efecto principal, estimamos de nuevo regresiones separadas para filmes dirigidos por mujeres y por hombres. En la submuestra de películas de directoras, los críticos varones no puntúan peor a las películas dirigidas por mujeres bajo ninguna escala a niveles convencionales de significatividad (p = 0,871 con estrellas enteras; p = 0,097 con medias estrellas). En la submuestra de películas de directores hombres tampoco se estiman diferencias en calificaciones bajo ninguna escala. Aunque el signo de los coeficientes en todos modelos muestra una penalización del trabajo directoral femenino, no se corrobora la moderación bajo la escala más fina observada en especificaciones anteriores.

Finalmente, estimamos una serie de modelos en los que la variable dependiente es de nuevo un indicador de si la película recibe 5 estrellas, y la variable dependiente clave vuelve a ser el género del crítico (cuadro n.º A2). El objetivo es comprobar si, como en la muestra completa, no hay un sesgo contra las directoras en el umbral más exigente de la escala, o si, como ocurría con las calificaciones medias, dicho sesgo se concentra en los críticos hombres.

En los modelos de efectos aleatorios, las críticas mujeres no penalizan a las directoras bajo ningún sistema. Los críticos hombres, en cambio, reducen entre 2,2 y 2,7 puntos porcentuales la probabilidad de otorgar cinco estrellas cuando califican bajo la escala con medias estrellas; no hay diferencia bajo la escala de estrellas completas. El efecto es modesto en magnitud, pero relevante en el extremo superior de la distribución. Al añadir efectos fijos de película, en cambio, los coeficientes pierden significación en todas las combinaciones, como ya sucedía en la muestra conjunta.

En resumen, los modelos son consistentes en su estimación de una penalización por parte de los críticos varones hacia las directoras, ya sea en la nota media o en la probabilidad de otorgar 5 estrellas, mientras que las críticas mujeres no presentan ningún patrón de sesgo. El papel moderador de la arquitectura de la escala no es estable. En las especificaciones con efectos aleatorios, la penalización se concentra bajo la escala más rígida al analizar la calificación media y bajo la escala con 9 puntos al analizar la máxima calificación. Al introducir efectos fijos de película, los coeficientes pierden significación y el patrón de escala se diluye.

Resultado 3: La penalización hacia las directoras está concentrada en los críticos hombres; la arquitectura de la escala modera este efecto, pero no de forma robusta. Las críticas mujeres no muestran ningún sesgo sistemático.

VI. DISCUSIÓN

Este artículo ha investigado cómo la arquitectura de las escalas de valoración condiciona la expresión de sesgos de género en la crítica cinematográfica. Partiendo de una muestra amplia de reseñas en medios españoles, hemos estudiado si la penalización hacia películas dirigidas por mujeres varía según la escala de calificación utiliza solo estrellas completas o permite medias estrellas, y si estos efectos dependen del género del crítico.

Los resultados muestran una penalización significativa para las directoras bajo la escala más rígida, que se atenúa y desaparece bajo la escala más fina. Esto confirma que la arquitectura de las evaluaciones no es neutral: las decisiones editoriales sobre el diseño de la escala influyen sobre el reconocimiento de ciertos grupos de creadores.

El análisis por género del crítico revela que la penalización es más visible entre los críticos varones. Bajo la escala de estrellas completas, estos otorgan sistemáticamente puntuaciones más bajas a las directoras, mientras que las críticas mujeres no muestran sesgo. Con medias estrellas, la penalización se reduce, lo que sugiere que la posibilidad de matizar las calificaciones limita la expresión de sesgos. Esto refuerza la importancia de considerar no solo quién evalúa, sino, también en qué condiciones institucionales se produce la evaluación.

En cuanto a la probabilidad de obtener la máxima puntuación, no hallamos evidencia clara de que las directoras se enfrenten a una desventaja. En la mayoría de las especificaciones, la brecha de género es pequeña y estadísticamente frágil. Esto sugiere que las mujeres no están sistemáticamente excluidas del reconocimiento simbólico más alto. Sin embargo, al introducir el género del crítico observamos de nuevo que los hombres tienden a conceder 5 estrellas a las directoras menos a menudo, en especial cuando disponen de medias estrellas. De nuevo, las críticas mujeres no muestran sesgo.

Estos resultados contrastan con dos predicciones que se desprenden de la literatura. Primero, contrastan con la hipótesis de que escalas más finas facilitan una discriminación sutil, al dar a los evaluadores margen para rebajar las calificaciones sin comprometer su (auto)imagen de imparcialidad (Rivera y Tilcsik, 2019; Botelho *et al.*, 2025). En la crítica cinematográfica, la posibilidad de otorgar medias estrellas parece que reduce, no amplifica, el sesgo contra las directoras. Este resultado es consistente con la idea de que el sesgo surge cuando el crítico debe calificar una película que percibe en un nivel intermedio y tiende a redondear hacia arriba si la película está dirigida por un hombre y hacia abajo si lo está por una mujer.

En segundo lugar, nuestros resultados no confirman que escalas con más categorías refuerzan la asociación cultural entre la nota máxima y atributos de excelencia, estereotípicamente masculinos, haciendo más difícil que las mujeres la reciban. Mientras que en el caso analizado por Rivera y Tilcsik (2019) el paso de un 10 a un 6 eliminó la brecha de género en la asignación de la máxima puntuación, en nuestro contexto esta se mantiene en 5 estrellas, preservando su valor simbólico. Esto puede explicar por qué no observamos que la escala empleada tenga un efecto sobre la penalización hacia las directoras en ese umbral de excelencia.

Dicho esto, nuestro estudio presenta varias limitaciones que conviene señalar. En primer lugar, la muestra se restringe a reseñas publicadas en cuatro medios digitales españoles, lo que limita la generalización de los resultados a otros contextos nacionales y culturales. En segundo lugar, solo contamos con dos cabeceras por tipo de escala, de modo que las diferencias observadas podrían estar influidas, pese al uso de efectos fijos de cabecera, por características idiosincráticas de cada medio o sus críticos más allá de la arquitectura de la puntuación. Por último, aunque la muestra total de reseñas es amplia, el número de películas dirigidas por mujeres es relativamente reducido, y lo es aún más cuando se cuentan las que entre ellas fueron reseñadas por críticas mujeres. Esto afecta a la potencia estadística y a la estabilidad de las estimaciones en algunas especificaciones.

Nuestros resultados tienen varias implicaciones. La primera es metodológica. Que observemos resultados diferentes, en ocasiones incluso opuestos, a los encontrados en la literatura previa, sugiere que

las conclusiones sobre los efectos de la arquitectura de escalas no son uniformes y dependen de cómo y dónde se investiguen. Esto subraya la necesidad de acumular evidencia en distintos contextos, a ser posible, incorporando indicadores externos de calidad, de manera que se pueda evaluar hasta qué punto los patrones observados responden a sesgos generalizables o a las especificidades de cada ámbito.

Por último, nuestros resultados sugieren que, además de promover una mayor diversidad entre críticos y creadores cinematográficos, repensar el diseño de los sistemas de puntuación puede ayudar a reducir sesgos en la evaluación del trabajo creativo. Estos sesgos no son un fenómeno inmutable ni homogéneo. Su manifestación depende de elementos institucionales aparentemente menores. Más allá del caso concreto de la crítica cinematográfica, invitamos a reflexionar sobre cómo la interacción entre arquitectura de evaluación y las características de los evaluadores puede afectar al reconocimiento del trabajo de mujeres y otros colectivos en distintos ámbitos culturales y profesionales.

NOTAS

(1) Estas estimaciones se realizaron a partir de modelos de efectos aleatorios cruzados que incluyen como componentes de varianza tanto el filme como el autor de la reseña.

(2) Los t-test correspondientes son t = –1,12, p = 0,262 para presupuesto, t = –0,35, p = 0,724 en calificaciones, t = –0,37, p = 0,714 en premios, y t = –0,20, p = 0,843 en nominaciones.

3) Una sencilla formalización ayuda a entender este punto. Sea q la probabilidad de que los críticos otorguen 5 estrellas a una película, p la probabilidad de que otorguen 4,5, r la de 4 y $s = p + r$. Asúmase que: los críticos penalizan a las películas dirigidas por mujeres solo en el rango de calificaciones (4,5), de tal forma que $sh > sm$, donde h y m denotan hombre y mujer, y, que la probabilidad de recibir 5 estrellas y la aportación a la media que proviene de las notas por debajo de 4 no dependen del género del director ni de la escala. En ese caso, la penalización de género bajo la escala de 5 puntos es $4(sm - sh)$ mientras que es igual a $4(sm - sh) + 0{,}5(pm - ph)$ bajo la escala de 9 puntos. Por tanto, si $sh > sm$ pero $ph < pm$ existe una brecha de género en la nota media que se atenúa bajo la escala de 9 puntos sin que la probabilidad de recibir 5 estrellas difiera por género.

BIBLIOGRAFÍA

Alan, S., Ertac, S., y Mumcu, A. (2018). Gender stereotypes in the classroom and effects on achievement. *Review of Economics and Statistics, 100*(5), 876-890.

Boring, A. (2017). Gender biases in student evaluations of teaching. *Journal of Public Economics, 145*, 27-41.

Botelho, T. L., Jun, S., Humes, D., y DeCelles, K. A. (2025). Scale dichotomization reduces customer racial discrimination and income inequality. *Nature, 639*, 395-403.

Brooke, S., y Rao, A. H. (2024). Designing for justice in freelancing: Testing platform interventions to minimise discrimination in online labour markets. *Big Data y Society*, January-March, 1-19.

Card, D., Dellavigna, S., Funk, P., y Iriberri, N. (2020). Are referees and editors gender neutral? *The Quarterly Journal of Economics, 135*(1), 269-318.

Card, D., Dellavigna, S., Funk, P., y Iriberri, N. (2022). Gender differences in peer recognition. *Econometrica, 90*(5), 1943-1969

Comscore. (2024). *2024 Year in Review: Definiendo el rumbo del 2025*. Edición España.

Cook, D. A., y Beckman, T. J. (2009). Does scale length matter? A comparison of nine- versus five-point rating scales for the mini-CEX. *Advances in Health Sciences Education, 14*, 655-664.

Cubel, M., y Sanchez-Pages, S. (2025). There Will Be Bias: Gender Patterns in Film Criticism. *Documento de Trabajo*.

Culpepper, D., White-Lewis, D., O'Meara, K., Templeton, L., y Anderson, J. (2023). Do Rubrics Live up to Their Promise? Examining How Rubrics Mitigate Bias in Faculty Hiring. *The Journal of Higher Education, 94*(7), 823-850.

Ductor, L., Goyal, S., y Prummer, A. (2023). Gender and collaboration. *The Review of Economics and Statistics, 105*(6), 1366-1378.

Dupas, P., Modestino, A. S., Niederle, M., y Wolfers, J. (2021). Gender and the dynamics of economics seminars. *NBER Working Paper,* 28494.

Eberhardt, M., Facchini, G., y Rueda, V. (2023). Gender differences in reference letters: Evidence from the economics job market. *The Economic Journal, 133*(655), 2676-2708.

Eutsler, J., y Lang, B. (2015). Rating Scales in Accounting Research: The Impact of Scale Points and Labels. *Behavioral Research in Accounting, 27*(2), 35-51.

Fitzpatrick, A., Beg, S., y Lucas, A. M. (2021). Gender bias in assessments of teacher performance. *AEA Papers and Proceedings, 111,* 190-195.

Goldin, C., y Rouse, C. (2000). Orchestrating impartiality: The impact of "blind" auditions on female musicians. *American Economic Review, 90*(4), 715-741.

Hengel, E. (2022). Publishing while female. *The Economic Journal, 132*(648), 2951-2987.

Hospido, L. (2020). Gender Gaps in the Evaluation of Research: Evidence from Submissions to Economics Conferences15. *Oxford Bulletin of Economics and Statistics, 82*(3), 591-614.

Lassen, I. M. S., Bizzoni, Y., Peura, T., Thomsen, M. R., y Nielbo, K. L. (2022). Reviewer preferences and gender disparities in aesthetic judgments. arXiv preprint arXiv:2206.08697.

Lassen, I. M. S., Moreira, P. F., Bizzoni, Y., Thomsen, M. R., y Nielbo, K. (2023). Persistence of gender asymmetries in book reviews within and across genres. En *Computational Humanities Research Conference 2023*. CEUR Workshop Proceedings.

Leslie, S.-J., Cimpian, A., Meyer, M., y Freeland, E. (2015). Expectations of brilliance underlie gender distributions across academic disciplines. *Science, 347*(6219), 262-265.

Loist, S., Verhoeven, D., Eikhof, D. R., Prommer, E., Ehrich, M. E., Jones, P., ... y Dadlani, A. (2024). *Re-framing the picture: An international comparative assessment of gender equity policies in the film sector.* Film university Babelsberg Konrad Wolf.

Muñoz, M. (2006). Legibilidad y variabilidad de los textos. *Boletín de Investigación Educacional, 21*, 13-25.

Nochlin, L. (1971). Why have there been no great women artists? *ArtNews, 69*(9), 22-39, 67-71.

Preston, C. C., y Colman, A. M. (2000). Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica, 104*, 1–15

Proudfoot, D., Kay, A. C., y Koval, C. Z. (2015). A gender bias in the attribution of creativity: Archival and experimental evidence for the perceived association between masculinity and creative thinking. *Psychological Science, 26*(11), 1751-1761.

Revilla, M. A., Saris, W. E., y Krosnick, J. A. (2014). Choosing the Number of Categories in Agree-Disagree Scales. *Sociological Methods y Research, 43*(1), 73-97.

Rivera, L. A. (2025). Two-point rating system cuts out racial bias. *Nature*, 639.

Rivera, L. A., y Tilcsik, A. (2019). Scaling Down Inequality: Rating Scales, Gender Bias, and the Architecture of Evaluation. *American Sociological Review, 84*(2), 248-274.

Samahita, M. (2024). Are economics conferences gender-neutral? Evidence from Ireland. *Oxford Bulletin of Economics and Statistics, 86*(1), 104-118.

Szigriszt Pazos, F. (1993). *Sistemas predictivos de legilibilidad del mensaje escrito: fórmula de perspicuidad*. PhD Thesis. Universidad Complutense de Madrid.

Weathers, D., Sharma, S., y Niedrich, R. W. (2005). The impact of the number of scale points, dispositional factors, and the status quo decision heuristic on scale reliability and response accuracy. *Journal of Business Research*, 58, 1516-1524.

Weijters, B., Cabooter, E., y Schillewaert, N. (2010). The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing*, 27, 236-247.

Yoon, G. (2024). No One Optimal Way to Measure People's Attitudes? Preferred Length of Scales in Advertising Research. *Journal of Current Issues y Research in Advertising, 45*(1), 43-70.

ANEXO

CUADRO N.º A1

ESTADÍSTICOS DESCRIPTIVOS DE RESEÑAS POR MEDIO

VARIABLE	*FOTOGRAMAS* (*n* = 1015)	*EAM* (*n* = 1012)	*CINEMANÍA* (*n* = 1400)	*LA RAZÓN* (*n* = 774)
Estrellas	3,53 [3,49, 3,57]	3,37 [3,32, 3,43]	3,40 [3,37, 3,44]	3,19 [3,13, 3,25]
Legibilidad	63,84 [63,51, 64,18]	59,71 [59,42, 60,00]	51,97 [51,71, 52,23]	50,98 [50,38, 51,57]
Simplicidad	57,29 [57,02, 57,55]	55,19 [54,98, 55,40]	32,57 [32,37, 32,78]	45,20 [44,71, 45,70]
Directoras	21,6% [19,2%, 24,2%]	27,9% [25,2%, 30,7%]	23,0% [20,9%, 25,3%]	22,9% [20,0%, 26,0%]
Críticos mujeres	21,6% [19,2%, 24,2%]	13,3% [11,4%, 15,6%]	28,5% [26,2%, 30,9%]	32,4% [29,2%, 35,8%]
Guionistas mujeres	26,2% [23,5%, 29,0%]	31,8% [29,0%, 34,7%]	28,0% [25,7%, 30,4%]	28,7% [25,6%, 32,0%]
Directoras de fotografía	12,0% [10,1%, 14,2%]	15,5% [13,4%, 17,9%]	13,2% [11,4%, 15,0%]	13,6% [11,3%, 16,1%]
Nota IMDb	6,42 [6,37, 6,46]	6,59 [6,54, 6,63]	6,35 [6,31, 6,40]	6,47 [6,41, 6,53]
Votos en IMDb	50,082 [43.503, 56.661]	36,088 [29.972, 42.204]	36,105 [31.330, 40.880]	44,531 [36.802, 52.259]
Taquilla	$108,52M [$88,44M, $128,61M]	$64,46M [$45,40M, $83,51M]	$87,66M [$70,56M, $104,76M]	$101,18M [$75,21M, $127,15M]
Presupuesto	$54.40M [$47.94M, $60.86M]	$39.39M [$32.55M, $46.23M]	$46.99M [$41.30M, $52.69M]	$47.61M [$39.75M, $55.47M]
Premios	10,54 [8,77, 12,31]	11,33 [9,59, 13,08]	8,04 [6,85, 9,23]	11,60 [9,58, 13,62]
Nominaciones	26,03 [22,59, 29,46]	26,88 [23,50, 30,25]	19,91 [17,42, 22,39]	28,68 [24,45, 32,92]
Drama	62,5% [59,4%, 65,4%]	72,3% [69,5%, 75,0%]	59,1% [56,5%, 61,6%]	66,7% [63,3%, 69,9%]
Comedia	30,1% [27,4%, 33,0%]	18,0% [15,7%, 20,5%]	28,1% [25,8%, 30,6%]	25,3% [22,4%, 28,5%]
Producción EE. UU.	39,3% [36,4%, 42,4%]	35,7% [32,8%, 38,7%]	33,7% [31,3%, 36,2%]	34,2% [31,0%, 37,7%]
Producción España	29,3% [26,5%, 32,1%]	14,4% [12,4%, 16,7%]	30,6% [28,3%, 33,1%]	23,0% [20,2%, 26,1%]

CUADRO N.º A2

PROBABILIDAD DE OBTENER 5 ESTRELLAS POR GÉNERO DEL DIRECTOR Y DEL CRÍTICO

VARIABLE	[1]	[2]	[3]	[4)]
Director mujer	-0,063 (0,045)	-0,063 (0,045)	-0,067 (0,049)	— —
Medias estrellas	-0,071* (0,037)	-0,069* (0,040)	-0,065 (0,042)	-0,043 (0,049)
Director mujer x medias estrellas	0,069 (0,045)	0,070 (0,045)	0,076 (0,050)	0,063 (0,055)
Crítico hombre	-0,031 (0,038)	-0,033 (0,038)	-0,027 (0,040)	-0,009 (0,045)
Crítico hombre x medias estrellas	0,047 (0,048)	0,046 (0,048)	0,052 (0,053)	0,036 (0,058)
Director mujer x medias estrellas x crítico hombre	0,047 (0,040)	0,050 (0,039)	0,048 (0,043)	0,022 (0,052)
Nominaciones	-0,075 (0,050)	-0,075 (0,050)	-0,087 (0,054)	-0,074 (0,063)
Premios	0,001*** (0,000)	0,001*** (0,000)	0,001*** (0,000)	— —
Log(votos IMDb)	-0,000 (0,000)	-0,000 (0,000)	-0,001 (0,000)	— —
Nota IMDb	0,015*** (0,003)	0,015*** (0,003)	0,016*** (0,004)	— —
Log(taquilla)	0,008 (0,005)	0,008 (0,005)	0,011* (0,007)	— —
EF cabecera	NO	SÍ	SÍ	SÍ
EF película	NO	NO	NO	SÍ
Observaciones	3.904	3.904	3.188	3.247

Notas: MLPs. Errores agrupados a nivel de crítico. Los modelos [1] a [3] incluyen variables de duración del filme, tipo de estreno, género, idioma y países de producción. * $p < 0,10$ ** $p < 0,05$ *** $p < 0,01$.